

Requested Patent: JP2000172614A

Title: INTERNET RETRIEVING APPARATUS ;

Abstracted Patent: JP2000172614 ;

Publication Date: 2000-06-23 ;

Inventor(s): YANAGIDA KEITON ;

Applicant(s): NEC CORP ;

Application Number: JP19980351931 19981210 ;

Priority Number(s): ;

IPC Classification: G06F13/00 ; G06F17/30 ;

Equivalents:

ABSTRACT:

PROBLEM TO BE SOLVED: To automatically retrieve and find out the uniform resource locator(URL) of a changed web page. SOLUTION: This apparatus 10 is composed of a web page retrieving part 11, link destination retrieving part 12, link destination contents reading part 13, read contents storage part 14, disk cache 15, disk cache internal contents comparing part 16 and compared result reporting part 17. A URL corresponding to the master of the web page desired to connect is retrieved, the URL of the link destination is searched from the retrieved web page, and the contents of the web page at the link destination are read and stored in the read contents storage part 14. Moreover, the contents before the web page desired to connect are compared with the contents of the link destination stored in the read content storage part 14 and with the rate of coincidence as a reference, it is assessed whether or not the contents of both the web pages are equal.

BEST AVAILABLE COPY

BEST AVAILABLE COPY

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2000-172614

(P2000-172614A)

(43) 公開日 平成12年6月23日 (2000.6.23)

(51) Int.Cl.

G 0 6 F 13/00  
17/30

識別記号

3 5 4

F I

G 0 6 F 13/00  
15/40

テーム (参考)

3 5 4 D 5 B 0 7 5  
3 1 0 C 5 B 0 8 9

審査請求 有 請求項の数 2 O L (全 4 頁)

(21) 出願番号 特願平10-351931

(22) 出願日 平成10年12月10日 (1998.12.10)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 柳田 直致

東京都港区芝五丁目7番1号 日本電気株式会社内

(74) 代理人 100108578

弁理士 高橋 留男 (外3名)

Fターム (参考) 5B075 K002 N002 N006 N035 NK10

P006 Q008

5B089 GA12 GA21 GB04 JA24 KA03

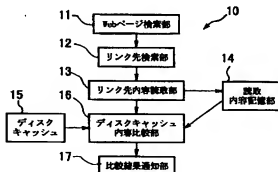
LB04 LB14 LB25

(54) 【発明の名称】 インターネット検索装置

(57) 【要約】

【課題】 変更されたWebページのURLを自動的に検索して見つける。

【解決手段】 インターネット検索装置10は、Webページ検索部11と、リンク先検索部12と、リンク先内容読取部13と、読取内容記憶部14と、ディスクキャッシュ15と、ディスクキャッシュ内容比較部16と、比較結果通知部17とから構成した。接続を希望するWebページの親に相当するURLを検索し、検索されたWebページからリンク先のURLを探し出し、リンク先のWebページの内容を読み取り、読取内容記憶部14に格納する。また、接続を希望するWebページの内容と、読取内容記憶部14に格納されたリンク先の内容とを比較し、両Webページの内容が同一であるか否かを一致率を基準として判断する。



## 【特許請求の範囲】

【請求項1】 World Wide Webにおいて、階層構造をなす複数のWebページから接続を希望するWebページを検索するインターネット検索装置であって、前記インターネット検索装置は、接続に失敗したWebページのURLよりも上位層のWebページのURLを推定して検索する手段と、前記上位層のWebページからリンク先を抽出する手段と、この抽出されたリンク先のWebページの内容を読み取って記憶する手段と、予め前記接続に失敗したWebページの以前の接続時における内容を記憶する手段と、この以前の接続時における内容と前記リンク先のWebページの内容とを比較して同一であるか否かを判断する手段とを備えていることを特徴とするインターネット検索装置。

【請求項2】 前記インターネット検索装置は、前記接続に失敗したWebページの以前の接続時における内容と、前記リンク先のWebページの内容とを比較する際に、両者の一致率が所定の一致率を超えたときに両者が同一のWebページであると判断し、前記両者の一致率が前記所定の一致率に満たないときに両者は異なるWebページであると判断する手段を備えていることを特徴とする請求項1に記載のインターネット検索装置。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明は、World Wide WebにおけるWebページを検索するインターネット検索装置に関する。

## 【0002】

【従来の技術】WWW (World Wide Web) はネットワーク上に分散された多数のHTTP (Hyper Text Transfer Protocol) サーバにより提供されるWebページが相互にリンクされたもので、WebページはHTML (Hyper Text Markup Languages) で記述されたテキストファイルと画像データ等から構成されている。クライアントとなるユーザはWebブラウザ等を使用することでHTTPサーバに接続し、WebページのデータをダウンロードしてWebページを表示する。各Webページはネットワーク上における所在がURL (Uniform Resource Locator、最近ではURI: Uniform Resource Identifierと呼ばれる。) によって指定されており、ユーザはWebブラウザ等において接続を希望するWebページのURLを入力することによって所望のWebページに接続することができる。ところで従来においては、例えばWebページの内容が何ら変更されていなくても、このWebページの内容が変更されて、移動先のURLが提示されない場合においては、ユーザはこのWebページに接続することができなくなり、移動先のURLを検索システム等を使用して探し出す必要が生じていた。

## 【0003】

【発明が解決しようとする課題】上記のように、URLが変更されたWebページを検索システム等を利用して検索する場合、一般に、検索用に入力するキーワードに対して複数のWebページの候補が存在するが、特に多量のWebページが候補として提示された場合には、目的とするWebページを見つけ出すまでに多くの時間と手間が必要になるという問題が生じる。また、接続を希望するWebページのURLが変更された直後においては、検索システム等に登録されている内容が更新されていないおそれがあり、移動先のURLを見つけ出すことが困難になるという問題が生じる。本発明は上記事情に鑑みてなされたもので、接続を希望するWebページのURLが変更された際にも、変更後のURLを自動的に検出することが可能なインターネット検索装置を提供することを目的とする。

## 【0004】

【課題を解決するための手段】上記課題を解決して係る目的を達成するために、請求項1に記載の本発明のインターネット検索装置は、World Wide Webにおいて、階層構造をなす複数のWebページから接続を希望するWebページを検索するインターネット検索装置であって、前記インターネット検索装置は、接続に失敗したWebページのURLよりも上位層のWebページのURLを推定して検索する手段と、前記上位層のWebページからリンク先を抽出する手段と、この抽出されたリンク先のWebページの内容を読み取って記憶する手段と、予め前記接続に失敗したWebページの以前の接続時における内容を記憶する手段と、この以前の接続時における内容と前記リンク先のWebページの内容とを比較して同一であるか否かを判断する手段とを備えていることを特徴としている。

【0005】上記構成のインターネット検索装置では、例えばHTTPサーバ側においてWebページの構成等を変更したことに伴って接続を希望するWebページのURLが変更され、クライアントとされるユーザから接続を希望するWebページへの接続ができなくなった際にも、接続を希望するWebページに比べてURLが変更される可能性の低い上位層のWebページに接続することによって、接続を希望するWebページに設定された新たなURLを探し出すことが可能である。

【0006】さらに、請求項2に記載のインターネット検索装置は、前記インターネット検索装置は、前記接続に失敗したWebページの以前の接続時における内容と、前記リンク先のWebページの内容とを比較する際に、両者の一致率が所定の一致率を超えたときに両者が同一のWebページであると判断し、前記両者の一致率が前記所定の一致率に満たないときに両者は異なるWebページであると判断する手段を備えていることを特徴としている。

【0007】上記のようなインターネット検索装置で

は、接続を希望するWebページのURLが変更されたことに加えて、接続を希望するWebページの内容が更新された場合であっても、この更新されたWebページの内容と、以前に接続した時点におけるWebページの内容との比較をおこなう際に、一致率を適宜に設定することによって、例えば両者の内容が完全に一致していなくても、両者が同一のWebページであると判断することができるため、接続を希望するWebページの変更後のURLを見つけることが可能である。

#### 【0008】

【発明の実施形態】以下、本発明のインターネット検索装置の実施形態について添付図面を参照しながら説明する。図1は本発明の一実施形態に係るインターネット検索装置を示すブロック構成図である。本実施形態によるインターネット検索装置10は、Webページ検索部11と、リンク先検索部12と、リンク先内容読取部13と、読取内容記憶部14と、ディスクキャッシュ15と、ディスクキャッシュ内容比較部16と、比較結果通知部17とから構成されている。Webページ検索部11は、例えば接続を希望するWebページの親(上位階)に相当するWebページのURL、すなわち例えば接続希望のURLがhttp://www.nec.co.jp/usr/index.htmlである場合に、このURLの親に相当するURLとして例えばhttp://www.nec.co.jp/index.html等のWebページを検索する。リンク先検索部12は、Webページ検索部11で検索されたWebページに対して、このWebページからリンクされているリンク先のWebページのURLを探し出す。リンク先内容読取部13は、リンク先のWebページの内容を読み取り、読取内容記憶部14に格納する。ディスクキャッシュ15は、予め、以前に接続した時点での接続希望のWebページの内容を記憶しておく。ディスクキャッシュ内容比較部16は、ディスクキャッシュ15に記憶されている接続を希望するWebページの以前の内容と、リンク先内容読取部13で読み込まれた読取内容記憶部14に格納されたリンク先のWebページの内容とを比較し、両Webページの内容が同一であるかを、予め設定された例えば一致率等を基準として判断する。比較結果通知部17は、ディスクキャッシュ内容比較部16での比較結果を通知する。

【0009】本実施形態によるインターネット検索装置10は上述の構成を備えており、次に、インターネット検索装置10の動作について図1および図2を参照しながら説明する。図2は図1に示すインターネット検索装置10の動作を示すフローチャートである。先ず、ユーザが例えばWebブラウザ等を使用して接続を希望するWebページのURLに接続を試みた時に、例えば該当するWebページが見つからない等のエラーがHTTPサーバ側から通知されると、インターネット検索装置10が起動する(ステップ1)。インターネット検索

装置10のWebページ検索部11は、接続に失敗したWebページのURLに基づいて、このWebページの親に相当するWebページのURLを推定して接続を行う(ステップ2)。ここで、例えば接続希望のURLがhttp://www.nec.co.jp/usr/index.htmlである場合には、このURLの親に相当するURLとして例えばhttp://www.nec.co.jp/index.htmlを採用する。

【0010】親に相当するWebページに接続すると、リンク先検索部12は、このWebページからリンクされているリンク先のWebページのURLを探して接続を行い(ステップ3)、リンク先内容読取部13がリンク先のWebページの内容を読み取って読取内容記憶部14に格納する(ステップ4)。一方、ディスクキャッシュ15には、接続を希望するWebページの以前の内容、すなわちURLが変更される前に接続した際に記憶されたWebページの内容が格納されている。そこで、ディスクキャッシュ内容比較部16は、ディスクキャッシュ15に記憶されている接続を希望するWebページの以前の内容と、リンク先内容読取部13で読み込まれた読取内容記憶部14に格納されたリンク先のWebページの内容とを比較し、所定の例えば50%以上の一致率が得られた際に両Webページの内容が同一であると判断し、一方、例えば50%未満の一致率では互いに異なるWebページであると判断する(ステップ5)。比較結果通知部17は、ディスクキャッシュ内容比較部16において両Webページが同一であると判断された際には、リンク先内容読取部13が読取したリンク先のWebページのURLをユーザに通知し、逆に両Webページが互いに異なるかと判断された際には、URLの検出に失敗した旨の通知をユーザに行う(ステップ6)。

【0011】本実施形態によるインターネット検索装置10によれば、例えばHTTPサーバ側においてWebページの構成等を変更したことに伴って接続を希望するWebページのURLが変更され、クライアントとされるユーザから接続を希望するWebページの接続ができなくなった際にも、接続を希望するWebページに比べてURLが変更される可能性が低い親のWebページに接続することによって、接続を希望するWebページに設定された新たなURLを探し出すことが可能である。また、例えばWebページの構成等に加えて、接続を希望するWebページの内容が更新された場合であっても、以前に接続した時点でのWebページの内容と、更新されたWebページの内容との比較の際に、適宜の一致率を基準値として設定することによって、両Webページの内容が完全に一致していなくても、両者が同一のWebページであると判断することができるため、接続を希望するWebページの変更後のURLを見つけることが可能である。

【0012】なお、本実施形態においては、接続に失

敗したWebページの親に相当するWebページのURLを推定としたが、これに限定されず、例えば同じHTTPサーバ上のURLであって、URLの末尾が例えばindex.htmlとなっているWebページを検索してもよい。ここで、例えばindex.htmlは、一般に、HTTPサーバの標準設定によって作成されるWebページであって、Webページの構成が変更される際にURLが消去される可能性が低い。要するに、接続の失敗したWebページのURLの近隣のURLであって、確実に接続が可能なURLを推定すればよい。

【0013】

【発明の効果】以上説明したように、請求項1記載の本発明のインターネット検索装置によれば、接続に失敗したWebページに比べてURLが変更される可能性の低い上位層のWebページに接続することによって、接続に失敗したWebページに設定された新たなURLを探し出すことが可能である。さらに、請求項2記載のインターネット検索装置によれば、接続を希望するWebページの内容が更新された場合であっても、以前に接続し

た時点でのWebページの内容と、更新されたWebページの内容との比較の際に、適宜の一致率を基準値として設定することによって、両Webページの内容が完全に一致していなくても、接続を希望するWebページの変更後のURLを見つけることが可能である。

【図面の簡単な説明】

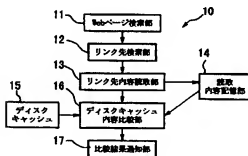
【図1】 本発明の一実施形態に係るインターネット検索装置のブロック構成図である。

【図2】 図1に示すインターネット検索装置の動作を示すフローチャートである。

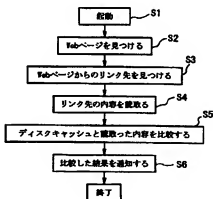
【符号の説明】

- 10 インターネット検索装置
- 11 Webページ検索部
- 12 リンク先検索部
- 13 リンク先内容読取部
- 14 読取内容記憶部
- 15 ディスクキャッシュ
- 16 ディスクキャッシュ内容比較部
- 17 比較結果通知部

【図1】



【図2】



**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

**BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☒ BLACK BORDERS

☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES

☐ FADED TEXT OR DRAWING

☐ BLURRED OR ILLEGIBLE TEXT OR DRAWING

☐ SKEWED/SLANTED IMAGES

☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS

☐ GRAY SCALE DOCUMENTS

☒ LINES OR MARKS ON ORIGINAL DOCUMENT

☒ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY

☐ OTHER: \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**